

Translation Difficulty Estimation

Gonzalo Cruz Garcia, Serge Sharoff
University of Leeds

Project overview

With increasing globalisation and the push towards automation, it becomes essential to improve on the current translation methods. In this project, the United Nations corpus is explored in order to understand the factors contributing to difficulties in both human and machine translation. The time taken for human translation of documents is predicted as well as classification of sentences based on their machine translation difficulty. This is achieved using a combination of lexical features, translation edit rate and sentence vectorisation through Facebook's XLM neural networks.

Data and methods

The aim of this study was to better inform decision making in the field of translation. This was achieved by predicting the difficulty of human translation and the usefulness of machine translation. A set of 300 timed documents were provided by the United Nations Office at Geneva. These contained the time at which translation started and ended. There is a strong linear correlation between the time taken to translate and the length of documents (figure 1), however, there is still great uncertainty.

The proportion of linguistic features, such as time adverbials or prepositions, can be used as an initial estimate of the difficulty of a text. Linear regression was used to predict the rate at which documents were being translated, based on these linguistic features. In order to improve on this method, Facebook's open-source cross-lingual language model (XLM) neural network was used to produce sentence embeddings. XLM takes sentences as input, outputting a vector based on an objective such as translation.

Due to the limited availability of timed data, Translation Edit Rate (TER) was used. TER is defined as the number of edits needed to make a machine translation match a reference human translation. TER scores range from 0 for a perfect translation with no edits required, to 1 where the entire sentence is changed. TER scores were computed for over 10 million sentences found in the UN-parallel corpus for both Spanish and French translations (figure 2). Sentences are fed through XLM and the output vectors are used as inputs for regression and classification (figure 3).

Due to the uncertainty when dealing with timed documents, a small subset of 300 timed sentences were produced (figure 4). These give us a closer understanding of the difficulty in human translation. A similar dataset produced by official UN translators will give better insight for specialised translation.

Key findings

Predicting translation rate at the document level yielded a correlation of 0.43 using linguistic features. When predicting TER at the sentence level, linguistic features only obtained a 0.17 correlation using linear regression, compared to 0.42 when using XLM. Classification of sentence based on a "bad / average / good" machine translation division obtained 53% f1-score when using XLM compared to 40% with linguistic features, both using a support vector machine.

Overall, linguistic features offer a good base estimate for prediction at a document level. On the other hand, sentence pre-training methods such as XLM offer a great improvement at the sentence level. Further information surrounding translation decisions, such as document priority or experience level of each translator, and the addition of fine-tuning is likely to generate a significant improvement in results.

Value of the research

This research established a first step towards collaboration between the University of Leeds and both the United Nations and World Trade Organisation.

The work done offers a baseline for predictions of translation times as well as usefulness of machine translations. These predictions can be used to improve current decision making and work allocation in the field of translation.

Research theme

- Natural Language Processing
- Human and machine translation
- Language representation pre-training

Partners

- United Nations
- World Trade Organisation
- The Alan Turing Institute
- Leeds Institute for Data Analytics

Translation Difficulty Estimation

Gonzalo Cruz Garcia, Serge Sharoff
University of Leeds

Images

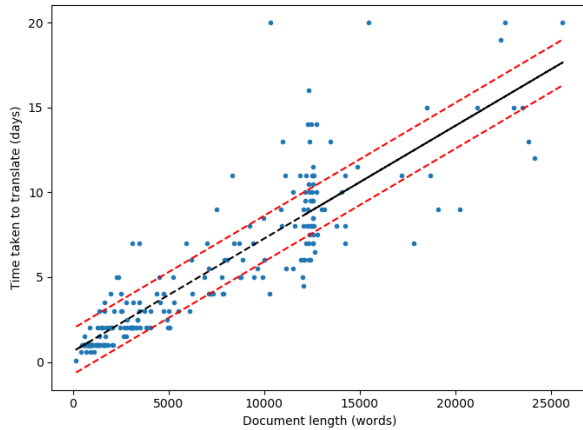


Figure 1. Time it took to translate a document against its length in words for a set of around 300 UN documents.

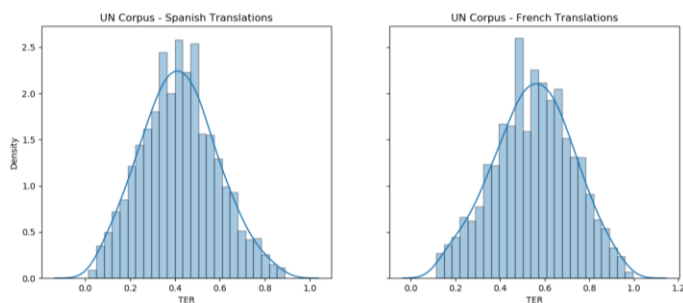


Figure 2. Distribution of machine translation scores (TER) for Spanish and French sentence translations in a sample of UN documents.

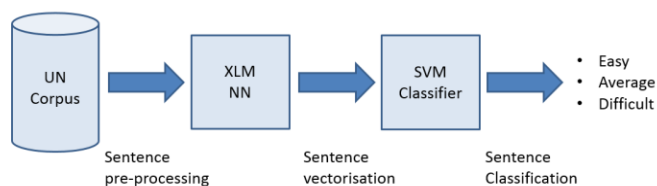


Figure 3. Example of the translation difficulty classification workflow used in this project.

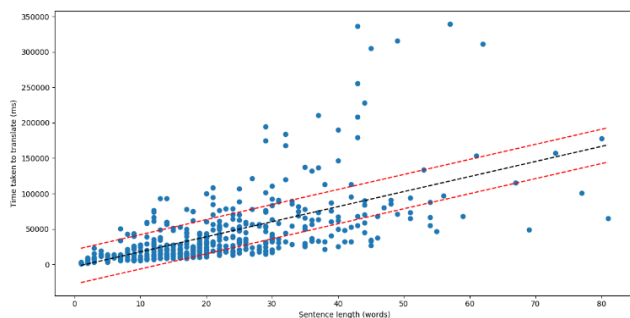


Figure 4. Time it took to translate a sentence against its length in words for a set of around 300 sentences. Sentences translated from English to either Spanish or French.